Nouvelles considérations pour la détection de réutilisation de texte

Fabien Poulard, Stergos Afantenos et Nicolas Hernandez
LINA (CNRS - UMR 6241)
2 rue de la Houssinière – B.P. 92208, 44322 NANTES Cedex 3
{prenom.nom}@univ-nantes.fr

Résumé. Dans cet article nous nous intéressons au problème de la détection de *réutilisation de texte*. Plus particulièrement, étant donné un document original et un ensemble de documents candidats — *thématiquement similaires* au premier — nous cherchons à classer ceux qui sont dérivés du document original et ceux qui ne le sont pas. Nous abordons le problème selon deux approches : dans la première, nous nous intéressons aux similarités *discursives* entre les documents, dans la seconde au *recouvrement de n-grams hapax*. Nous présentons le résultat d'expérimentations menées sur un corpus de presse francophone construit dans le cadre du projet ANR PIITHIE.

Abstract. In this article we are interested in the problem of *text reuse*. More specifically, given an original document and a set of candidate documents — which are *thematically similar* to the first one — we are interested in classifying them into those that have been derived from the original document and those that are not. We are approaching the problem in two ways : firstly we are interested in the *discourse similarities* between the documents, and secondly we are interested in the *overlap of n-grams that are hapax*. We are presenting the results of the experiments that we have performed on a corpus constituted from articles of the French press which has been created in the context of the PIITHIE project funded by the French National Agency for Research (*Agence National de la Recherche, ANR*).

Mots-clés : réutilisation de texte, recouvrement de n-grams hapax, similarités discursives, corpus journalistique francophone.

Keywords: text reuse, hapax n-grams overlap, discourse similarities, french journalistic corpus.

1 Introduction

« La réutilisation de texte est l'activité par laquelle des textes écrits pré-existants sont réutilisés pour créer de nouveaux textes ou versions [...] il y a réutilisation quand il y a une réalisation consciente d'une transformation d'un texte pour en arriver à un autre » (Clough & Gaizauskas, 2008). La duplication (copie à l'identique), la révision, l'adaptation de genre, le résumé, la traduction, la citation,... sont autant de formes différentes de réutilisation d'un texte original. Les chercheurs comme les industriels sont conscients depuis de nombreuses années de l'intérêt d'étudier cette activité qui correspond à des enjeux applicatifs réels : la détection de documents dupliqués sur le web a des conséquences sur l'efficacité des moteurs du recherche aussi bien pour leur traitement (coût d'indexation et de stockage) que sur la précision des réponses ramenées. La détection de plagiats présente aussi un grand intérêt pour le respect du droit d'auteur que cela concerne le code source de logiciels ou tout document servant "de base" à des devoirs d'étudiants par exemples. Le suivi d'impact d'une communication sur un produit ou sur une information rendue publique présente aussi des intérêts commerciaux et scientifiques dans une perspective de veille.

En pratique, les systèmes de détection de réutilisation de texte procèdent selon trois étapes :

1. d'abord ils sélectionnent des types d'unités textuelles à observer (mot, syntagme, phrase, paragraphe, document, n-gram avec/sans recouvrement) dans les documents manipulés ;
2. ensuite ils construisent un modèle de chaque document par normalisation linguistique (lexicale, syntaxique, sémantique) ou numérique (condensation par algorithme de hashage) et par filtrage (les mots pleins, un n-gram donné, les n-premiers rencontrés dans le texte, pondérés par $tf.idf$, ...) des observables ;
3. enfin ils comparent effectivement les documents sur la base de ces représentations.

Le choix de la représentation est bien entendu dépendant de la méthode de comparaison utilisée. Celles-ci varient suivant différents coûts de traitement : de mesures de similarités rencontrées en Classification ou en Recherche d'Information (RI) (ratio des matériaux partagés, distance vectorielle) aux comparaisons plus complexes et spécifiques (plus longues sous chaînes communes, distance d'édition) (Uzuner *et al.*, 2004; Metzler *et al.*, 2005; Seo & Croft, 2008; Bendersky & Croft, 2009; Clough & Gaizauskas, 2008).

Les différentes étapes de cette procédure sont sujètes à de nombreux enjeux techniques : comment choisir les unités textuelles les plus représentatives du contenu du document ? Les moins coûteuses à extraire (en terme de ressources requises, de méthodes à mettre en place, de temps de calcul) ? Les plus caractéristiques des phénomènes de réutilisation ? Quelles sont les méthodes les plus adéquates en termes de précision et de temps de traitement pour détecter une forme donnée de réutilisation ?...

Le contexte applicatif du présent article est celui de la détection de réutilisations à partir d'un écrit original, dans des textes journalistiques francophones présentant des similarités thématiques avec le document source[1]. En particulier notre tâche consistait à classer les documents candidats comme étant des réutilisations ou non d'un document original connu. La figure 1 fournit un exemple issu de notre corpus d'un texte original, d'une réutilisation de celui-ci et d'une similarité thématique (cas de non réutilisation).

Dans cette article, nous proposons de nouvelles considérations théoriques afin de mieux cadrer

[1]Ce travail a bénéficié du soutien de l'Agence Nationale de la Recherche, projet PIITHIE (www.piithie.com) portant la référence 2006 TLOG 013 03.

Exemple (1) Texte original :
Le groupe Carrefour va prochainement se lancer dans la VOD en France mais pas seulement. Fier de ses parts de marché dans la vente de DVD dans d'autres pays d'europe (autour de 13%), Carrefour va aussi ouvrir son service de Vidéo à la Demande en Espagne, Italie et Belgique.

Exemple (2) Réutilisation :
Le géant national de la grande distribution française lancera une offre de VOD, vidéo à la demande en France, Belgique, Italie et Espagne, où par-ailleurs il détient une part de marché de 13,3% dans la vente de DVD.

Exemple (3) Similarité thématique :
Le 8 novembre 2006 en partenariat avec Orange, Carrefour lancera son offre de téléphonie mobile : Carrefour Mobile. Le groupe de distribution devient ainsi opérateur virtuel de téléphonie mobile (MVNO) avec les mêmes ambitions que son concurrent direct Auchan.

FIG. 1 – Classification de documents candidats comme étant des réutilisations ou non d'un document original connu

le problème de la détection de réutilisation de textes. Nous introduisons notamment la notion de *singularités* d'un document vis-à-vis d'une collection. Nous avançons que cette considération permet de sélectionner plus finement les unités textuelles représentant un document. Cela offre de nouvelles perspectives telles que : l'indexation des documents du web sans filtrage des réutilisations par post-traitement, l'exploitation des moteurs de recherche traditionnels pour la recherche directe de réutilisations, la récupération de cas de réutilisations avec transformations importantes, la distinction des documents thématiquement similaires de ceux qui constituent effectivement des réutilisations.

L'étude de cette propriété est envisagée autour de deux expériences de détection de réutilisation de texte. Celles-ci utilisent des unités textuelles jusqu'alors non considérées dans la littérature pour représenter les documents : des *marques discursives* et des n*-grams hapax*[2]. Nous observons leur capacité de détection de réutilisations de par leur singularité. En pratique, elles seront utilisées en complément avec d'autres observables tels que des termes sélectionnés sur leur $tf.idf$ afin de capturer par les similarités thématiques inter-documents.

1.1 Cadre théorique et terminologie

Les caractéristiques d'une réutilisation de texte doivent remplir deux tâches : d'une part définir si un document candidat est une réutilisation et d'autre part déterminer de quel document il est une réutilisation. Par caractéristiques, nous entendons toute marque linguistique telle que les termes, la syntaxe des phrases, ou des marques de plus haut niveau telles que les références à des entités ou l'organisation des idées.

La première tâche nécessite de considérer les caractéristiques que l'on retrouve à la fois dans le document dérivé et dans les documents originaux, nous les appelons : les *invariants*. Il n'est pas possible de sélectionner les invariants pour un document original si on n'a pas défini à quel document candidat on le comparait. Nous parlerons de classes d'invariants pour désigner des

[2]Hapax signifie « qui a été dit qu'une fois ».

invariants de différentes natures linguistiques, sans nous limiter aux formes de surface. Ainsi, une référence à une même entité entre deux documents à l'aide de formes de surfaces différentes est considérée comme un invariant ; l'invariant étant la référence à ladite entité, et non la forme employée pour la dénommée.

La seconde tâche nécessite de considérer les caractéristiques présentes uniquement dans le document et absentes de la collection homogène à laquelle celui-ci appartient, nous les appelons : les *singularités*. Cette collection homogène se définit par des critères spécifiques sélectionnés selon la finalité désirée. Il n'est pas possible de calculer les singularités d'un document sans avoir auparavant défini la collection de documents dans laquelle il s'inscrivait. Les singularités peuvent prendre des formes aussi diverses que les invariants et de la même façon nous parlerons de classes de singularités pour désigner les singularités de différentes natures linguistiques.

Une combinaison de marques peut correspondre à une singularité ou un invariant même si ce n'est pas le cas des marques prises individuellement. Nous essayons d'estimer dans cet article dans quelle proportion certaines classes de singularités sont généralement invariantes et permettent donc d'identifier des réutilisations.

2 Etat de l'art

Notre énonciation du principe de singularité est appuyée par divers travaux de la littérature.

La prise en compte de l'importance d'un terme dans une collection est un principe considéré dès les premiers travaux en RI (Jones, 1972; Salton & Buckley, 1988). Ainsi la bien connue inverse de la fréquence des documents idf est un facteur utilisé avec la fréquence des termes d'un document tf pour relativiser cette dernière mais aussi pour favoriser les termes présents dans peu de document d'une collection. Elle se calcule en générale en prenant le logarithme[3] du ratio du nombre de documents de la collection sur le nombre de documents distincts dans lesquels on retrouve un terme donné.

Pour la tâche de reconnaissance du style d'un auteur, (van Halteren, 2004) montre que le « comptage de [combinaison de] traits linguistiques d'un texte normalisé par sa longueur et leur déviation par rapport à la moyenne observée sur un corpus de référence » permet d'obtenir de meilleurs résultats que les méthodes fondées sur des analyses en composante principale, des analyses discriminantes linéaires, ou des distributions probabilistes.

Afin de mesurer l'appartenance d'un terme au genre d'une collection dans laquelle il apparaît, (Hernandez, 2004) a utilisé avec succès une des composante de l'idf, le nombre de documents distincts dans lesquels on retrouve un terme, pour filtrer les termes des documents.

Dans le contexte de la détection de réutilisation le long d'un spectre de degrés de similarités, (Metzler *et al.*, 2005) observent qu'une mesure fondée sur le recoupement de mots pondérés par idf est l'une des deux méthodes les plus performantes pour la détection de réutilisation avec fortes transformations (reprise partielle des faits). Suivant les méthodes considérés, leurs taux de précision varie de 40–60% lorsque la similarité traduit un lien thématique ou qu'il s'agit d'une reprise partielle de faits, à 80–100% pour les reprises identités.

[3] Car le simple ratio donne des valeurs très grandes.

3 Méthodes de détection de réutilisation

Nous présentons ci-après deux méthodes se fondant sur le cadre théorique que nous avons défini. La première utilise des marques discursives et la seconde des *hapax*.

3.1 À l'aide de marques discursives singulières

Nous décrivons dans cette section une approche basée sur la modélisation des documents par une représentation partielle de leur structure discursive. Nous pensons que cette structure est suffisamment singulière (*cf. section 1.1*) pour permettre de détecter automatiquement les reprises globales de documents. Nous détaillons ci-dessous la réflexion qui nous a mené à cette hypothèse puis nous en présentons une modélisation afin de mener nos expérimentations.

La restructuration discursive globale d'un document est une tâche difficile qui nécessite une réécriture partielle de ce dernier et une réorganisation des idées. Dans le cadre des articles de presse, il s'agit de retravailler la structure argumentative ou descriptive de l'article original. Nous choisissons de nous intéresser aux éléments de structuration discursifs car nous supposons qu'ils varient peu lors d'une reprise globale d'un document.

La structure discursive est complexe à extraire et à caractériser, excepté lorsqu'elle est marquée par des connecteurs discursifs non ambigus (Sporleder & Lascarides, 2008). Nous choisissons alors plus particulièrement de travailler à partir des connecteurs discursifs et de ne considérer uniquement la structure discursive des documents rendue visible par ces connecteurs. Nous sommes tout à fait conscient que ces derniers ne sont que la surface émergée de l'iceberg du discours. Toutefois, ce racourci nous permet d'expérimenter l'idée de la détection automatique de reprise par comparaison des structures discursives sans nécessiter d'appliquer des méthodes complexes à mettre en œuvre. En effet, les connecteurs discursifs appartiennent à une classe lexicale aisément observable par des techniques automatiques, peu ambigue (Sporleder & Lascarides, 2008) et dont l'utilisation peu fréquente dans les documents est en accord avec le principe de singularité énoncé précédemment.

Nous choisissons de représenter la structuve discursive sans tenir compte de l'ordre d'apparition desdits connecteurs et en considérant ces derniers dans leur forme lexicale observée. Nous sommes conscients de la naïveté de cette approche, la division en classe sémantique, la position relative ou absolue et la séquence d'apparition sont des éléments à prendre en compte. Cependant, dans le cadre d'une reprise globale du document, notre approche fonctionne assez bien comme le montre les résultats de la section 4.3 et la simplicité de l'approche est un atout pour sa mise en œuvre.

Chaque document est modélisé par un vecteur du nombre absolu des occurrences de connecteurs discursifs. Ces connecteurs recherchés sont au nombre de 90 et ont été collectés ou traduits de la littérature (Knott, 1996; Marcu, 1997) dans (Hernandez, 2004). Il s'agit principalement d'adverbes tels que *premièrement*, *ensuite*, ... ou des groupes adverbiaux tels que *dans un premier temps*, *tout d'abord*, ...

Nous comparons ensuite le vecteur d'un document considéré original et celui d'un document candidat à l'aide d'une mesure de similarité cosinus. Les valeurs varient de 0 à 1 où 0 signifie que les deux vecteurs sont indépendants et 1 signifie que les vecteurs sont identiques. Nous choisissons de retourner 0 lorsqu'aucun connecteur n'apparaît dans un des deux documents, ce

qui n'est pas entièrement satisfaisant étant donné que 22% des documents de notre corpus ont cette caractéristique.

En résumé, nous cherchons à définir si une réutilisation réuni deux documents. Nous comparons pour cela les structures discursives rendues visibles par les connecteurs de ces documents. Nous expérimentons cette méthode dans la section 4.3.

3.2 À l'aide de marques lexicales singulières

Nous posons ici l'hypothèse que des hapax ont un pouvoir discriminant pour la détection de réutilisation. Nous décrivons dans cette section notre processus pour obtenir des hapax ainsi que notre technique de comparaison de documents pour détecter des réutilisations à partir de ces hapax.

Dans le cadre de ce travail, nous avons choisi d'observer les n-grams hapax comme instance des hapax. Pour ce faire nous avons produit des unigrams, bigrams, trigrams, ..., sept-grams à partir de notre corpus en filtrant les mots vides. Nous avons ensuite cherché à identifier les n-grams hapax de chaque document en sommant pour chaque n-gram son nombre d'occurrences dans le document et son nombre d'occurrences dans les documents des autres répertoires de notre corpus ; un répertoire réunie un document source original et un ensemble de documents candidats dérivés ou non du document source. Seuls les n-grams qui apparaissent une seul fois ont été conservés (hapax).

En ce qui concerne la méthode de comparaison nous avons abordé le problème comme une tâche de classification binaire où il s'agissait de classer un document candidat en document dérivé ou en document non-dérivé. Pour chaque paire de documents original et candidat, nous avons constitué un vecteur de deux attributs : le nombre de hapax *en commun* et le nombre des hapax distincts.

4 Expérimentations

Dans cette section, nous décrivons le corpus utilisé pour nos expérimentations lesquelles sont présentées à la suite.

4.1 Construction et composition du corpus PIITHIE

Le corpus utilisé lors des expérimentations a été construit dans le cadre du projet ANR PIITHIE[4]. Nous présentons ci-dessous sa construction et sa composition.

Dans un premier temps, des documents récents considérés comme originaux ont été manuellement sélectionnés sur des sites de presse en ligne. Un article était considéré comme original s'il avait pour source une agence de presse (AFP, REUTERS). Dans une période postérieure immédiate, des documents dérivés candidats ont été récupérés à l'aide de moteurs de recherche sur des sites sélectionnés du web. La sélection des sites de recherche visait à garantir une homogénéité de genre des documents ramenés. Les requêtes des moteurs étaient produites à partir

[4]Le corpus produit sera prochainement distribué sur `http://www.piithie.com`.

Classifieur	Classe	Précision	Rappel	F-mesure
App. Référence	D_R	0.86	0.44	0.58
	$D_{\bar{R}}$	0.48	0.88	0.62
Connecteurs	D_R	0.94	0.56	0.70
	$D_{\bar{R}}$	0.56	0.94	0.70
Hapax	D_R	0.923	0.895	0.909
	$D_{\bar{R}}$	0.83	0.873	0.851

TAB. 1 – Comparaison des résultats des différentes expérimentations

des cinq mots les plus « rares » des documents originaux. Le degré de rareté d'un mot était calculé a priori sur la base du nombre de documents qu'une requête avec ce mot ramenait dans les moteurs de recherche utilisés. Les documents récupérés ont ensuite été annotés manuellement par deux annotateurs afin classer les documents comme dérivés ou non-dérivés. Un document était considéré comme dérivé s'il reprenait les évènements décrits dans le document source et s'il présentait des sous-chaînes de mots communes.

Au total, le corpus se compose de 77 documents originaux, 496 documents dérivés et 293 documents non-dérivés. En pratique, le corpus est divisé en 77 répertoires contenant chacun un document original et un ensemble de documents candidats (dérivés et non-dérivés).

4.2 Approche de référence

Dans cette section, nous décrivons une méthode dont nous utiliserons les résultats comme référence pour nos expérimentations. Cette méthode naïve reprend les principes de la méthode discursive présentée à la section 3.1, la différence réside dans le choix des observables. En effet, nous n'utilisons pas ici la structure discursive mais un sous-ensemble du lexique du corpus pour caractériser les documents.

Nous utilisons un lexique de 90 mots extraits aléatoirement du corpus et comparable en taille avec la signature discursive, en filtrant les mots outils à l'aide d'une heuristique. Chaque document est alors caractérisé par le nombre d'occurrences des éléménts de ce lexique qui y apparaît. Les vecteurs obtenus sont alors comparés à l'aide d'une mesure cosinus et l'on classe comme repris les documents qui présentent une valeur supérieure à un seuil fixé. Inversement les documents obtenant une valeur inférieure sont considérés comme des non-reprises. Le seuil a été sélectionné de manière à maximiser la précision et le rappel de cette méthode sur le corpus.

Le tableau 1 expose les résultats de l'approche de référence pour la classe des documents repris (D_R) et celle des documents non-repris ($D_{\bar{R}}$). Outre l'obtention de résultats pour comparaison, la similarité de cette approche avec la méthode discursive permet de tester le rôle particulier des connecteurs dans la section suivante.

4.3 Expérimentations sur les connecteurs discursifs

Nous expérimentons dans cette section la détection automatique de reprise sur le corpus présenté à la section 4.1 par la méthode discursive décrite à la section 3.1.

Nous avons observé au préalable que les formes graphiques des connecteurs de notre dictionnaire sont présentes dans environ 80% des documents. Certaines de ces formes sont ambigues mais nous n'en tenons pas compte (*cf. section 3.1*). Nous avons considéré chaque document original du corpus Piithie, et pour chacun de ces documents nous avons calculé le cosinus de son vecteur et celui de chacun des documents candidats associés.

Nous avons comparé les valeurs du cosinus entre les originaux et les candidats en différenciant les dérivés des non-dérivés. Nous observons que ces valeurs pour les non-dérivés sont réparties de manière assez homogène sur tout l'espace image alors qu'elles se densifient autour de 1 (vecteurs identiques) pour les dérivés. Cette observation supporte notre hypothèse que les structures discursives sont proches entre les documents originaux et les dérivés et qu'il s'agit donc d'un invariant potentiel.

Étant donné un document original et sa modélisation selon notre méthode, nous classons comme repris un document candidat dont le cosinus appliqué entre son modèle et celui de l'original est supérieur à un seuil fixé. Par opposition, les documents pour lesquels le cosinus est inférieur au même seuil sont considérés comme des non-dérivés. Nous avons choisi le seuil de 0.8 car c'est la valeur du cosinus qui maximise la précision du classifieur sur le corpus. Nous avons défini cette valeur en faisant varier le cosinus par dizième. Le classifieur obtient alors une précision de 94% et un rappel de 56% comme le montre le tableau 1.

La caractérisation des documents par un vecteur d'occurrence des connecteurs donne des résultats meilleurs que l'approche de référence pour les deux classes, autant en terme de précision que de rappel, comme le montre le tableau 1. Ainsi il est de 8 points supérieurs en précision et 12 points supérieurs en rappel pour la classe des documents dérivés. Ces résultats semblent supporter notre hypothèse d'autant plus que la seule variation avec l'approche de référence provient des observables (lexiques aléatoire vs. connecteurs discursifs). Les meilleurs résultats confirment le rôle particulièrement discriminant des connecteurs.

En résumé, nous avons rapproché originaux et dérivés en nous basant sur la distribution des connecteurs discursifs. L'expérience montre que les documents qui sont des dérivés d'un original conservent globalement les connecteurs de ce dernier. À l'opposé, les documents qui sont des faux positifs partagent aléatoirement ces connecteurs. Ceci supporte notre hypothèse que la structuration du discours permet de distinguer les dérivés parmi les documents candidats.

4.4 Expérimentations sur les n-grams hapax

Nous rapportons ici nos expérimentations de classification de documents en dérivés et non-dérivés d'un document original selon l'approche décrite à la section 3.2.

Pour ce faire, nous avons utilisé trois algorithmes de classification différents[5] : *Naïve Bayes*, *Sequential Minimal Optimization for training support vector machines (SMO)* et *LogitBoost*. Pour ce dernier algorithme, nous avons fait varier le nombre d'itérations de *boosting*.

Nous avons conduit une évaluation par validation croisée à dix partitions. Nos résultats sont présentés dans le tableau 1. Comme ce tableau l'illustre (troisième ligne : « Hapax »), la meilleure F-measure est obtenue avec *LogitBoost* en utilisant 15 itérations à la fois pour l'identification de documents dérivés D_R et non-dérivés $D_{\bar{R}}$. Néanmoins globalement les différents algorithmes

[5]Nous avons utilisés les algorithmes sus-mentionnés au sein de la plate-forme d'apprentissage automatique WEKA (Witten & Frank, 2005).

donnent de bons résultats compris entre 85% et 90% ce qui dépasse l'approche de référence définie à la section 4.2 de 3 points.

5 Discussion et perspectives

Dans cet article, nous avons proposé de nouvelles bases théoriques pour cadrer le problème de détection de réutilisation textuelle. Nous avons ainsi défini deux notions capitales, l'invariance et la singularité, dont la considération permet d'envisager autrement les étapes de la procédure de détection (par exemple le degré de singularité d'un trait d'un document constitue un nouveau critère de sélection pour représenter ce document). Nous montrons notamment que des hapax ou des marques singulières de nature discursive sont des indices probants pour différencier un document dérivé d'un document non-dérivé à partir d'un document original. Nous avons observé que les connecteurs discursifs et les n-grams hapax, singularités des documents originaux, se retrouvaient dans les documents dérivés, ce qui nous a permis de les repérer. L'utilisation des connecteurs discursifs singuliers et des n-grams hapax constitue en soi une originalité de ce travail puisque ces marques n'avaient jusqu'à présent pas été considérées dans la littérature pour la détection de réutilisation de texte.

Les résultats élevés que nous obtenons sur notre corpus nous conduisent à vouloir confronter nos méthodes à d'autres données. En perspective à ce travail, nous projetons de réitérer nos expériences sur le corpus anglais METER (Clough, 2003; Clough & Gaizauskas, 2008).

Remerciements

Nous tenons à remercier nos relecteurs pour leurs critiques constructives et pour leurs suggestions d'amélioration.

Références

BENDERSKY M. & CROFT W. B. (2009). Finding text reuse on the web. In *WSDM '09 : Proceedings of the Second ACM International Conference on Web Search and Data Mining*, p. 262–271, New York, NY, USA : ACM.

CLOUGH P. & GAIZAUSKAS R. (2008). Corpora and text re-use. In A. LÜDELING & M. KYTÖ, Eds., *Corpus Linguistics : An International Handbook*, Handbücher zur Sprache und Kommunikationswissenschaft/Handbooks of Linguistics and Communication Science, chapter 59. Berlin : Mouton de Gruyter.

CLOUGH P. D. (2003). *Measuring Text Reuse*. PhD thesis, University of Sheffield.

HERNANDEZ N. (2004). *Description et Détection Automatique de Structures de Texte*. PhD thesis, Université Paris-Sud XI.

JONES K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, **28**, 11–21.

KNOTT A. (1996). *A Data-Driven Methodology for Motivating a Set of Coherence Relations*. PhD thesis, Department of Artificial Intelligence, University of Edinburgh.

MARCU D. (1997). *The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts*. PhD thesis.

METZLER D., BERNSTEIN Y., CROFT W. B., MOFFAT A. & ZOBEL J. (2005). Similarity measures for tracking information flow. In *CIKM '05 : Proceedings of the 14th ACM international conference on Information and knowledge management*, p. 517–524, New York, NY, USA : ACM.

SALTON G. & BUCKLEY C. (1988). Term-weighting approaches in automatic text retrieval. In *Information Processing and Management*, p. 513–523.

SEO J. & CROFT W. B. (2008). Local text reuse detection. In *SIGIR '08 : Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, p. 571–578, New York, NY, USA : ACM.

SPORLEDER C. & LASCARIDES A. (2008). Using automatically labelled examples to classify rhetorical relations : An assessment. **14**(3), 369—416.

UZUNER O., DAVIS A. & KATZ B. (2004). Using empirical methods for evaluating expression and content similarity. In *In 37th Hawaiian International Conference on System Sciences (HICSS-37). IEEE Computer Society*.

VAN HALTEREN H. (2004). Linguistic profiling for author recognition and verification. In *ACL '04 : Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, p. 199, Morristown, NJ, USA : Association for Computational Linguistics.

WITTEN I. H. & FRANK E. (2005). *Data Mining : Practical Machine Learning Tools and Techniques*. San Francisco : Morgan Kaufmann, second edition.